CORPUS

Corpus

2 | 2003
La distance intertextuelle

s. n. – Matemáticas y Tratamiento de Corpus. Logroño : Fundacion San Millán de la Cogolla, 2002, 350 p.

Sylvie Mellet

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/43
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 décembre 2003
ISSN : 1638-9808

Référence électronique
Sylvie Mellet, « *s. n. – Matemáticas y Tratamiento de Corpus*. Logroño : Fundacion San Millán de la Cogolla, 2002, 350 p. », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 02 mai 2019. URL : http://journals.openedition.org/corpus/43

s. n. — Matemáticas y Tratamiento de Corpus. Logroño : Fundacion San Millán de la Cogolla, 2002, 350 p.

Sylvie Mellet

1 Ce recueil rassemble les textes des communications faites au deuxième séminaire de la « Escuela interlatina de altos Estudios en Lingüística aplicada », qui s'est tenu à San Millán de la Cogolla en septembre 2000 et dont le thème, en cette année mondiale des mathématiques, était précisément : *Matemáticas y Tratamiento de Corpus*.

2 Après un prologue assez général de A. M. Municio, les vingt exposés sont répartis en quatre chapitres correspondant aux différentes sessions du séminaire :

- 1. Quantification et formalisation dans l'étude linguistique des corpus. La linguistique de corpus et ses paramètres – Identification par les méthodes statistiques du genre des textes et des styles individuels.
- 2. Etablissement automatique de modèles et de règles à partir de corpus, grâce à la modélisation mathématique de l'usage de la langue.
- 3. Méthodes automatiques et semi automatiques pour l'annotation des corpus textuels aux divers niveaux des analyses demandées par les études linguistiques et les humanités.
- 4. De la quantité à la qualité : techniques mathématiques pour classifier, visualiser et évaluer les données philologiques et culturelles. Nouvelles tendances pour l'usage et la conservation de l'héritage culturel. (Cette dernière partie traitant de questions situées aux marges de la linguistique[1], nous rendrons compte uniquement des trois premières).

3 Le chapitre I regroupe donc des articles qui, chacun à leur façon, associent peu ou prou au traitement automatique des corpus des traitements quantitatifs ; divers outils statistiques sont sollicités pour l'analyse des textes et leur comparaison à des fins varié (linguistiques, d'analyse du discours, d'extraction de l'information, de diffusion ciblée).

4 É. BRUNET est chargé de l'exposé introductif : « Formalisation et quantification des textes. Le domaine français » (pp. 15-34). Il y présente quelques bases textuelles en ligne (*Intratext*, *Gallica*, *Frantext*), les principaux outils d'annotation disponibles pour qui veut

lemmatiser et étiqueter un corpus textuel français, puis il expose les derniers développements de son logiciel *Hyperbase* dans la version qui permet justement de traiter des étiquettes grammaticales et d'intégrer ainsi des données morpho-syntaxiques aux analyses quantitatives réalisées par ce logiciel. Des exemples de traitements appliqués d'une part au théâtre classique, d'autre part à trois traductions françaises des quatre Evangiles permettent d'évaluer les avantages de cette nouvelle version, mais aussi la robustesse des résultats sur les grandes divisions du corpus et les regroupements de textes – que l'on compare de manière interne le traitement par *Hyperbase* des catégories grammaticales, des lemmes et des formes orthographiques, ou que l'on confronte de manière externe ces résultats à ceux obtenus avec l'aide d'autres logiciels tels que *Alceste* ou *Sphinx*.

5 B. PINCEMIN s'intéresse aussi aux regroupements automatiques de textes, mais dans une autre perspective : « Similarités texte-textes. Expérience d'une application de diffusion ciblée et propositions » (pp. 35-52). Après une rapide évocation des diverses applications possibles du calcul des similarités textuelles, B. P. présente donc différentes méthodes de calcul dérivées du modèle de l'espace vectoriel ; elle rappelle les principes et formules d'une demi-douzaine de calculs de distances, en en soulignant les avantages, les inconvénients et les principes récurrents de fonctionnement ; elle relève aussi les unités d'analyse qui pourraient y être intégrées, mais qui sont généralement négligées. Au terme de cette évaluation critique très pertinente et très enrichissante pour le lecteur, B. P. propose une « adaptation du modèle de l'espace vectoriel à la textualité » (p. 47 *et sq*.) qui prendrait en compte le contexte des unités textuelles, leurs interactions réciproques et la force du lien qui peut unir certaines d'entre elles. Cette contribution est particulièrement claire et stimulante.

6 A. SALEM présente un nouvel outil logiciel (implémenté dans la dernière version de son logiciel d'analyse lexicométrique *Lexico3*), qui permet de visualiser la localisation de toutes les occurrences d'une unité (ou d'un « type ») dans le corpus étudié : « Topographie textuelle dans l'analyse quantitative des textes »(pp. 53-59). La notion de « type » recouvre une nouvelle unité lexicométrique constituée d'un « ensemble d'occurrences sélectionnées parmi les occurrences [constitutives] du texte » (p. 55) ; un tel ensemble peut être défini en compréhension ou en extension et ses concrétisations les plus familières sont les champs sémantiques (voire les réseaux thématiques) et les familles lexicales. Quant à la visualisation de la répartition de ces types dans le corpus, elle se fait au moyen d'une représentation spatiale des textes sous forme de succession linéaire de petits carrés dont chacun représente un paragraphe du texte ; chaque ligne du graphe figure 50 paragraphes consécutifs et se trouve donc formée de 50 petits carrés. Ceux-ci sont coloriés d'une teinte plus ou moins intense selon qu'ils contiennent plus ou moins d'occurrences du type recherché. La visualisation est donc immédiate des lieux dans lesquels le type est particulièrement abondant ; divers repères permettant d'indiquer sur le graphe la structure du texte ou les étapes de la série chronologique étudiée aident alors à interpréter cette topographie textuelle.

7 D. MALRIEU et F. RASTIER approfondissent ici un thème qu'ils ont développé dans diverses autres publications[2] : « Genres et variations morphosyntaxiques » (pp. 61-84). Le principe méthodologique de leur étude consiste à former un vaste corpus hétérogène de textes français, d'époques variées et relevant de types de discours et de genres différents, à étiqueter ces textes au moyen de l'analyseur *Cordial*, à les classer *a priori* dans des discours, champs génériques, genres et sous-genres prédéfinis, puis à vérifier au moyen

de divers calculs de fréquences portant sur les catégories morphosyntaxiques si cette classification intuitive mais culturellement sanctionnée est confirmée par des corrélations significatives. Trois méthodes de calcul sont utilisées : l'analyse univariée, l'analyse en composantes principales et la classification hiérarchique ascendante.. Tous les résultats sont positifs en ce qu'ils manifestent très clairement la diversité des normes linguistiques attachées aux différents discours et genres et confirment la classification préalable. On remarquera que les auteurs se refusent à pratiquer une classification endogène « aveugle » à la Biber (qui pourrait avoir sa place soit en phase exploratoire, soit en phase de confirmation des résultats) au motif que « les résultats de la classification n'auraient pas pu être interprétés sans un patient travail de classification des genres, champs génériques et discours » (p. 81). Cependant, on peut alors se demander si une telle méthode est falsifiable et si elle ne conduit pas à toujours retrouver les informations externes qu'on a injectées dans les données dès le début de la procédure. On est tenté aussi de formuler quelques réserves sur la constitution du corpus (en dépit de l'argumentaire du § 2.1) : quel est l'intérêt, pour ce type de recherches, de superposer aux variations génériques de fortes variations chronologiques ? En d'autres termes, pourquoi faire figurer dans ce corpus échantillonné des textes du XVIème, XVIIème et même XVIIIème siècles ? certes ils sont en nombre restreint ; mais ils perturbent quand même l'analyse. Les récits ont d'ailleurs droit à une subdivision chronologique à laquelle les autres sous-genres n'accèdent pas. On peut encore se demander quel est l'impact des déséquilibres numériques au sein des diverses classes : ainsi, au premier niveau hiérarchique, le discours scientifique est représenté par 66 textes, le discours littéraires par 2074 ; le genre narratif, on regroupe 49 contes, 51 nouvelles et 989 romans « sérieux ». Enfin, aucune distinction n'est faite entre théâtre en prose et théâtre en vers ; le poids de ce dernier dans le genre en question serait pourtant intéressant à connaître, notamment pour interpréter la classification hiérarchique ascendante (p. 76) qui regroupe très rapidement le théâtre et la poésie (G3 et G4).

8 J. DENDIEN, à partir de son expérience au service de l'exploitation de *Frantext*, propose « Une théorie des objets textuels et des moteurs de recherche dans les bases textuelles » (pp. 85-93). Il montre d'abord les insuffisances de la plupart des moteurs de recherche actuels travaillant en « plein texte », dont les uns, très performants, sont fort lents et les autres, très rapides, n'ont que des capacités limitées ne permettant aucune des recherches complexes qui intéressent généralement les stylisticiens ou les linguistes. Or le logiciel *STELLA*, développé par J. D. à Nancy pour l'exploitation de *Frantext*, prouve que les deux qualités sont pourtant compatibles ; et ce sont donc les principes de fonctionnement qui ont présidé à sa réalisation que J. D. expose et analyse ici. On retiendra deux points qui nous paraissent fondamentaux et caractéristiques du logiciel *STELLA* : la réflexion sur la notion d'objet textuel qui, partant des « objets textuels atomiques » (p. 87) proches des unités linguistiques élémentaires qui font l'objet des requêtes de base, construit ensuite, au moyen de quelques lois de composition récurrentes et grâce à la constitution de listes, des « objets textuels complexes » ou « composites » ; en guise d'illustration est décomposée de manière très simple la requête unique qui extrait toutes les occurrences des syntagmes <(*maison* ou *cabane*) (*rouge foncé* ou *blanche*)> ; l'intégration de quelques « objets textuels natifs » (p. 89) tels qu'un fantôme ou un joker de longueur fixe ou variable permet encore d'accroître aisément les performances du moteur de recherche. Le second point remarquable consiste à associer à cette théorie des objets textuels la possibilité pour l'utilisateur « d'écrire ses requêtes de manière progressive, hiérarchisée, réutilisable et modulaire » (p. 92) : ce sont les « hyper-

grammaires », recueil de règles pouvant se référencer l'une l'autre ; elles offrent, selon l'auteur, la « possibilité de définir les recherches les plus complexes avec une précision 'chirurgicale' » (p. 93).

9 La deuxième session du séminaire commence par un long article de P. ALLEGRINI et V. PIRRELLI : « Entropia e modelli del linguaggio » (pp. 97-125), qui se signale par sa clarté d'exposition et ses qualités pédagogiques : on y retrouvera en effet, patiemment exposées, les définitions précises de notions fondamentales en la matière telles que les chaînes de Markov, les automates d'états finis, les processus stochastiques, l'entropie dite de Shannon, etc., le tout illustré d'exemples imagés et éclairants. Le revers de la médaille est sans doute que la progression de l'exposé paraît lente et que le lecteur a parfois l'impression d'être face à un cours plutôt qu'à un article scientifique ; néanmoins mieux vaut cet excès de prudence didactique que le défaut inverse. L'apport original de cette contribution est un effort pour intégrer le rôle de l'entropie dans l'analyse mathématique du langage et de fournir par là-même un cadre théorique fort permettant de transcender l'opposition entre deux approches antérieures jusqu'ici divergentes : celle qui mettait en évidence la force de corrélations syntaxiques à courte portée indépendantes de la nature des textes étudiés et celle qui, à l'inverse, dégageait la variabilité des corrélations plus globales selon les types de textes. Il semblerait que, de dichotomiques, ces approches puissent devenir complémentaires dans le cadre d'une analyse de l'entropie conçue non plus seulement comme une mesure de la probabilité qu'a une forme linguistique d'apparaître à tel ou tel moment du texte, mais comme une « forme moyenne de la distribution des événements linguistiques au fil du texte, de leurs relations et de leurs dépendances » (p. 123).

10 F. YVON, sous le titre « Classification approaches for linguistic analysis » (pp. 127-143)[3], fournit un récapitulatif des méthodes de classification inspirées par les technologies du *machine learning* (« apprentissage machine ») et les illustre par quelques applications possibles à des problèmes linguistiques. Le plan de l'exposé suit l'opposition traditionnelle entre approches symboliques et méthodes statistiques. L'évaluation proposée est succincte et reprend les arguments habituels de manque de couverture pour les premières qui ne demeurent robustes que dans des domaines langagiers restreints et de manque de pouvoir explicatif des secondes. Cette dernière critique nous paraît d'ailleurs peu pertinente : ou bien on reste dans le domaine du TALN et l'essentiel est alors l'efficacité ; ou bien on passe à des applications en linguistique ou en analyse du discours et il est alors sain que l'outil informatique redonne la main au scientifique pour lui laisser le soin de l'interprétation.

11 C'est aussi un récapitulatif que nous propose L. LEBART dans un article consacré à l'« Analyse multi-dimensionnelle de textes [avec] application aux questions ouvertes dans les enquêtes » (pp. 145-152).On trouve exposées les principales méthodes d'analyse multidimensionnelles avec quelques perfectionnements récemment apportés aux outils de base. L'attention prêtée aux spécificités du traitement des réponses aux questions ouvertes d'enquêtes permet à L. L. d'évoquer le problème posé par la nécessaire intégration –dans ce contexte – de méta-informations (méta-données linguistiques et informations extérieures sociolinguistiques, chronologiques, etc.). Elle le conduit aussi à signaler les caractéristiques des tableaux de données issus de ce type de corpus : pluralité des regroupements possibles, existence de matrices de profils lexicaux extrêmement creuses ; dans ce dernier cas, l'introduction de filtres sémantiques (soit extraits *a priori* de

la méta-information disponible, soit construits à partir du texte lui-même) devient indispensable pour améliorer les calculs de similarités.

12 Enfin, pour clore cette deuxième partie, l'article d'H. RODRÍGUEZ « Adquisición automática y uso de taxonomías de amplia cobertura » (pp. 153-170) s'intéresse aux taxinomies à large couverture indépendantes de l'ontologie propre à un domaine particulier. Une fois encore, c'est un état des lieux qui est présenté, avec un rappel des différents types d'ontologie et des principes qui ont présidé à leur construction ; puis vient une description succincte de quelques ontologies lexico-conceptuelles à large couverture : *the Unified Medical Language System* (*UMLS*), *WordNet*, *EuroWordNet* et leurs différentes utilisations possibles. L'article s'achève par une partie plus technique consacrée aux méthodes de construction automatique de telles ontologies.

13 Le troisième chapitre de l'ouvrage s'ouvre sur un article de J.-P. CHANOD : « Eléments de méthodologie pour une analyse syntaxique robuste » (pp. 173-183). L'auteur souligne la faiblesse des analyseurs syntaxiques à base de règles et en donne un exemple très probant avec les phénomènes d'accord (loi d'unification) : un analyseur qui tient pour acquis qu'en français le verbe et le sujet s'accordent en nombre et en personne ou que les déterminants d'un substantif s'accordent avec celui-ci en genre et en nombre semble répercuter une règle élémentaire et consensuelle de la syntaxe française. Pourtant nombreux sont les faits de langue qui contreviennent à cette règle, en dehors même de toute négligence ou erreur grammaticale. Puis l'auteur évoque les propriétés fondamentales d'un bon analyseur, au premier rang desquelles l'incrémentalité qui s'accompagne de l'autonomie descriptive de chaque opération et donc aussi de la fragmentation descriptive des phénomènes complexes. On voit alors combien le cadre théorique pour la construction de tels analyseurs syntaxiques s'écarte des approches formelles de la syntaxe qui ont nourri les diverses théories linguistiques et, en conclusion, l'auteur suggère qu'à ce titre de tels analyseurs syntaxiques robustes soient aussi utilisés comme instruments d'expériences de falsification permettant de délimiter la champ de validité d'un principe linguistique.

14 Viennent ensuite deux exemples concrets de construction d'un analyseur syntaxique en lien avec la constitution d'un corpus arboré. Le premier concerne le français et est présenté par A. ABEILLÉ, L. CLÉMENT, A. KINYON et F. TOUSSENEL : « Un corpus français arboré : quelques interrogations » (pp. 185-195). On regrette la rédaction hâtive de cet article qui compte un certain nombre de formulations maladroites, de redites, d'enchaînements logiques abrupts et qui ne prend guère la peine justifier les choix faits, en matière d'étiquetage notamment ; si bien qu'on en vient à penser qu'on aura là, au terme du travail, un jeu d'étiquettes de plus, parmi tous ceux qui existent déjà, et dont l'utilité et la réutilisabilité restent douteuses dans la mesure où il n'est confronté à aucun autre et que la question des standards n'est pas posée. Le parseur est développé sur le mode incrémental et les premiers résultats de son exploitation linguistique confirment les acquis de la linguistique théorique et de la psycho-linguistique.

15 Le deuxième exemple concret est beaucoup plus intéressant ; il porte sur l'italien puisqu'il s'agit d'une description détaillée et scrupuleuse des principes et méthodes de construction de la « Treebank sintattico-semantica dell'italiano di Si-Tal » (pp. 221-243 ; Si-Tal = Sistema Integrato per il Tratamento Automatico del Linguaggio). Cette présentation est faite par S. MONTEMAGNI et A. LENCI. On note immédiatement le souci des auteurs de situer leur travail dans un cadre international, de le confronter avec les méthodes qui ont présidé à la création d'autres corpus arbors, de tenir compte des

standards prônés dans le cadre de grands programmes européens, de justifier tous leurs choix d'annotations spécifiques. Les annotations se distribuent sur cinq niveaux : orthographique et métatextuel, morpho-syntaxique, syntaxique de constituants, syntaxique fonctionnel et lexico-sémantique. L'un des points-clés de cette structuration, sur lequel les auteurs insistent beaucoup, est la séparation du niveau d'annotation syntaxique en deux niveaux indépendants bien que complémentaires : on a là deux « axes de saisie orthogonale » d'un même niveau de représentation, ce qui offre de multiples avantages, entre autres celui de rendre le système d'annotation plus indépendant de toute théorie linguistique sous-jacente et celui de régler élégamment et simplement le problème des positions vides, traces et autres indices qui compliquent généralement les représentations syntaxiques arborées ou parenthésées.

16 Entre ces deux articles consacrés à la constitution de corpus arborés et au développement d'analyseur syntaxique, que nous avons donc rapprochés pour faciliter notre synthèse, prennent place une note de quatre pages d'I. CORTAZAR et L. HERNANDEZ consacrée à « Modelado de lenguaje en reconocimiento de habla » (pp. 197-201) et un article de J. G. PEREIRA, V. ROCIO, M. F. XAVIER et G. VICENTE : « Criação automática de uma colecção de textos de português medieval parcialmente anotados sintacticamente » (pp. 203-220) où les outils développés pour étiqueter et parser un corpus de portugais médiéval s'appuient sur les points de ressemblance, apparemment assez nombreux, entre cet état ancien de la langue et son état moderne.

17 Enfin la session III s'achève par un article de P. ALLEGRINI, A. LENCI, S. MONTEMAGNI et V. PIRRELLI : Le forme del significato. Acquisizione e rappresentazione dell'informazione semantica » (pp. 245-268). L'approche sémantique est ici une approche contextuelle dans laquelle le sens des mots s'appréhende à travers leurs interactions dynamiques (on s'étonne que, se plaçant dans une telle perspective, les auteurs aient l'air d'ignorer les travaux de C. Fuchs et B. Victorri). Après l'examen de quelques implications théoriques et pratiques de cette conception, les auteurs décrivent un procédé d'acquisition automatique fondé sur la similarité dsitributionnelle des mots dans un contexte préannoté (là encore on regrette qu'aucun des chercheurs français ayant recours à ce type de méthode, finalement assez banal, ne soit cité en bibliographie).

18 Soulignons, pour terminer, un trait dominant de ce recueil : il présente davantage un état de l'art que des pistes de recherche vraiment nouvelles. A une ou deux exceptions près, les articles qui le composent offrent tous une longue partie récapitulative de bilan ou de présentation des outils existants ; les suggestions d'applications qui suivent sont le plus souvent données à titre d'exemples et peu approfondies. Néanmoins un tel état des lieux n'est sans doute pas inutile dans le paysage foisonnant et relativement dispersé des traitements de corpus. En outre, la plupart des auteurs se sont appliqués à dégager des principes méthodologiques plutôt qu'à simplement décrire le fonctionnement des outils. Leurs synthèses ouvrent donc la porte au dialogue et à la réflexion critique et posent les premiers éléments d'une évaluation de ce champ disciplinaire en plein renouveau.

19 Signalons enfin que la présentation matérielle de l'ouvrage est élégante, mais que les coquilles dans les textes sont assez nombreuses.

NOTES

1. Voici les articles qui la composent : Bozzi A. « Nuove tendenze per la conservazione e l'utilizzo del patrimonio librario nell'era digitale » ; Fedele G. « Restauro di documenti a stampa antichi per il riconoscimento automatico dei caratteri » ; Broia D. « Algoritmi e scienze umanistiche : il digitale per il recupero della conoscenza » ; Belcastro C. & Eisinberg A. « Cluster analysis per l'attribuzione di paternità in corpora testuali » ; Benel A. & Calabretto S. « 'Exploration' de corpus de documents archéologiques à l'aide de théories algébriques ».

2. Cf. notamment Malrieu D. & Rastier F. (2001). « Genres et variations morpho-syntaxiques », *TAL* 42, 2 : 547-577

3. Personnellement, nous regrettons que l'auteur ait cru nécessaire de rédiger son texte en anglais dans le cadre d'une publication de « l'Escuela interlatina » et d'un séminaire où, visiblement, l'intercompréhansion passive entre langues romanes fonctionnait fort bien.